From:	POULSEN Mike
To:	Eric Blischke/R10/USEPA/US@EPA
Cc:	PETERSON Jenn L; ANDERSON Jim M
Subject:	Portland Harbor SQG discussion issues
Date:	12/10/2010 05:27 PM

Eric,

To prepare for discussion on Monday, here are DEQ's comments on the LWG recent benthic approach submittal. Let us know if you would like to discuss before the meeting Monday afternoon. If you want to come to DEQ for the conference call, we have Room E reserved for the afternoon.

- Mike

Floating Point Model, DRAFT documentation of FPM Development:
Criteria for development of an initial list versus a final list are outlined in the memo. The chemical list is critical in developing an appropriate and relevant SQGs. However, there are several steps in this process that currently need to be revised before a final model is submitted.

1. Statistical Difference Between Hit / No Hit Distributions: The use of parametric methods (ANOVA) have been shown to be inaccurate in determinations of significant difference between distributions and must be revised in the final FPM. The use of non-parametric methods is required for comparison of hit and no hit distributions for the initial list of chemicals. DEQ's review has found that these distributions are often non-normal and the variances are not equal. When assumptions of normality are violated the power the distinguish between distributions is reduced, introducing a bias towards not finding a difference between distributions. It is unclear what is meant by a "non-parametric" t-test (ANOVA) since these tests assume normality. DEQ recommends either Wilcoxon Mann Whitney test (which can be done in using EPA's Pro UCL software) or permutation tests which can test the difference between means without relying on a distributional form and variance. Chemicals found to have a difference using the non-parametric tests should be included in both the initial list and final lists.

2. Chemical List by Species and Endpoint: The determination of statistical significance and associated chemical list should be species and test specific. These lists should not be expected to be the same between endpoints and species. Currently, the same chemical list is used for all species and endpoints (a determination of statistical significance for a chemical in one test endpoint means inclusion for all species and tests endpoints). If the chemical list included for each species and endpoint is not correlated with observed toxicity, this could impact the evaluation of reliability for a given set of SQGs developed for the model.

3. Removal From Chemical List: Chemicals should not be removed from the list of final SQGs based on criteria that "removal caused no change in any of the overall error and reliability rates". Overall reliability is not the only measure of interest. It appears the criteria for removing chemicals the final list was the absence of false positives for a given chemical, or

those set at an AET (the highest no hit above which everything is a hit). However, chemicals set at this level does not also mean there aren't hits below this level that shouldn't be evaluated or that inclusion of the chemical did not contribute to false negatives. If chemicals are removed, the model must be re-optimized to show there are no associated differences in the chemical concentration of the SQGs. Otherwise, inclusion of these chemicals in the model without using the associated values may alter the chemical concentration SQGs for other chemicals and associated reliability.

4. Use of SQGs for Conventional Chemicals (%OC and % fines): If conventional parameters are included in the model, then the associated SQGs that must be used as clean-up goals. These parameters may be acting as surrogates for chemistry correlated with toxicity not included in the model. For example, levels of organic carbon and % fines at levels greater than what would occur naturally are found co-located with highly contaminated areas. If conventionals are not to be used as SQGs (as indicated in this memo), then they should be removed from the model. This will allow the chemistry to correlate with toxicity instead. Inclusion of these parameters in the model alters the chemical concentration SQGs for other chemicals and associated reliability.

5. Chemistry Data: It appears that some chemicals were excluded if they were designated as "non-CERCLA". It should be clarified which chemicals were removed on this basis. Again, all chemicals should be evaluated for inclusion based on an appropriate test between hit and no hit distributions. For example, it is not clear if or how TPH-gasoline, TPH diesel, or TPH residual or TPH fractions were evaluated in the model.

6. Page 4, Relationship to Toxicity Criterion: Distributions for each chemical should be compared using non-parametric tests. It is unclear why Test/Control <1 is specified in this step of model development in relation to relationship with toxicity. Samples that had Test/Control $\geq$ 1 should be included in model runs, just designated as "no statistical difference" as defined comparison on whether the test response is significantly less than the control response. Please clarify.

7. Documentation: Documentation should include output from statistical tests between hit and no hit distributions for all chemicals, and all spreadsheets related to FPM model development.

Reliability of PECs and PELs and PEC / PEL Quotient Reliability:

The reliability of PECs, PEL and mean quotients were submitted as an Excel spreadsheet, but the list of PECs and PELs and the methodology for calculating the quotients was not clear. Additional information should show a chemical list and associated PEC or PEL for each chemical detected and the list used in the calculation of the mean quotient. Currently it is not clear if this list is based on available PECs and PELs, or a reduced list.

Quantification of Uncertainty in Portland Harbor Bioassay Response:

We have not fully reviewed the LWG's evaluation, so we are not ready to accept it. We acknowledge

that there will be uncertainty associated with sample results. It would be helpful to see if the replicate data are normally distributed (a key assumption for using the normal distribution function).

Note that for an evaluation of L2 (such as in the development of a predictive model) we are really interested in whether the sample is L2 or L3. In other words, we want to know if a sample is L2 or worse; we are less interested in whether it is exactly L2. On Figure 4, we would look at the addition of the bars for red (correctly predicted at L2) and green (under predicted – should be L3). This gives us the likelihood that we have accurately predicted toxicity of L2 or greater. These additive likelihoods appear to go from 50% up to a little less than 90% (assuming we accept their evaluation of probabilities).

Randomization Tests:

We are not clear on the purpose of the evaluation using randomization. We already have information to determine the ability of a model to predict toxicity greater than chance. Chance would be the expected outcome given the dataset. For a hit in any bioassay, the rate is about 22% hits, 78% no-hits. If a model results in reliabilities of predicted hits = 39% and predicted no-hits = 92%, then we know this is an improvement over chance (39% probability of hit if we exceed screening values versus 22% probability of hit without any information; 92% probability of no-hit if we are below screening values versus 78% probability of no-hit without any information).

The proposed randomization approach appears to test the model reliability versus another model based on randomized data. We do not understand the value of this comparison.

-